ISOLATED SPEECH RECOGNITION USING ARTIFICIAL NEURAL NETWORKS

Prasad D Polur[1], Ruobing Zhou[1], Jun Yang[1], Fedra Adnani[1], Rosalyn S. Hobson[2]
[1]Department of Biomedical Engineering, Virginia Commonwealth University, Virginia, USA,
[2]Department of Electrical Engineering, Virginia Commonwealth University, Virginia, USA

Abstract-**In this project Artificial Neural Networks are used as research tool to accomplish Automated Speech Recognition of normal speech. A small size vocabulary containing the words YES and NO is chosen. Spectral features using cepstral analysis are extracted per frame and imported to a feed-forward neural network which uses a backpropagation with momentum training algorithm. The network is trained to recognize and classify the incoming words into the respective categories. The output from the neural network is loaded into a pattern search function, which matches the input sequence with a set of target word patterns. The level of variability in input speech patterns limits the vocabulary and affects the reliability of the network. The results from the first stage of this work are satisfactory and thus the application of artificial neural networks in conjunction with cepstral analysis in isolated word recognition holds promise.**

I. INTRODUCTION

The goal of this research is to develop a human machine interface (HMI) the can recognize and classify spoken words under normal conditions and speech that is impaired or distorted in some manner. In this study the researchers will develop an accurate speech recognition system, which is capable of dealing with distinct speech inputs, such as distorted speech (Dysarthric Speech), stressed or elongated speech, rather than just normal speech. Once this system is fully developed it can aid end users who are disabled or those who are in medical distress, i.e intensive care unit hospital patients. In the first stage of the research the authors developed a HMI that uses mel cepstrum coefficients and artificial neural networks (ANNs) to perform speaker dependent word identification. The authors chose these two technologies because limited work has been done combining these two applications for speech processing and these technologies were developed using biological systems as their model. This paper presents the results from the first stage of this research.

II. METHODOLOGY

The general system architecture is discussed in the following sections. The steps involved in the Process of Speech Recognition are as follows:

- Sampling and Digitizing the speech
- Spectral Feature extraction
- Classification of time frames as phoneme categories
- Matching the category scores obtained to target (word) scores
- Determining the best match of the scores/patterns for identification of the word, i.e. word mapping

In this study speech is directly obtained in digitized form using the TASCAM DAP1 Digital Audio Tape Recorder. The speech is sampled and recorded at 44.1-kHz and directly loaded into the computer. The speech is then down sampled to 11 kHz (reduced storage requirement and ease of processing) and preprocessed using a digital audio editing software called SOUND FORGE 4.5 (noise and distortion reduction).

System Architecture

The overall architecture of this speech recognition is shown in figure one.

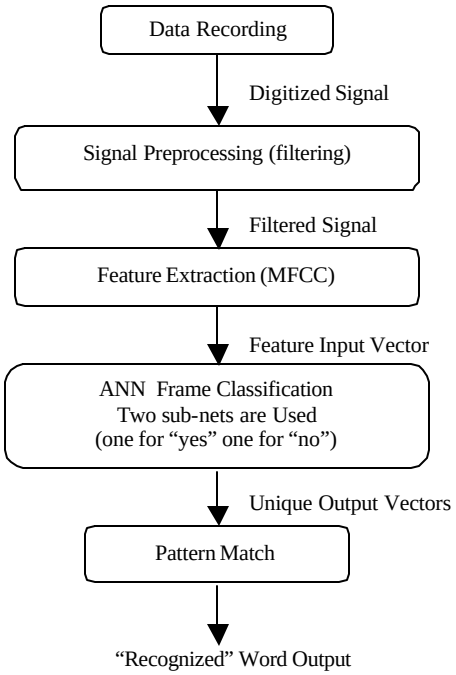

Figure 1. System Architecture

Feature Extraction (Mel Frequency Cepstrum Coefficients)

After obtaining the spectrogram of the speech waveform (see figure 2), we proceed to extract the spectral features, for each individual frame. The digitized

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
25OCT2001	N/A	-

Title and Subtitle Isolated Speech Recognition Using Artificial Neural Networks	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Department of Biomedical Engineering, Virginia Commonwealth University, Virginia	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)

Distribution/Availability Statement
Approved for public release, distribution unlimited

Supplementary Notes
Papers from the 23rd Annual International Conference of the IEEE ENgineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom., The original document contains color images.

Abstract

Subject Terms

Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU

Number of Pages
4

waveform is converted into a spectral-domain representation. We extract twelve Mel-frequency cepstral coefficients (MFCC coefficients) and one log energy feature (for a total of 13 features per frame) from the spectrum of the speech. Feature extraction based on Mel Frequency Cepstral Coefficients (C_i) uses a bank of twenty filters with center frequencies and bandwidths determined by the Mel model proposed in [1]. The MFCC's are calculated using:

$$C_i = \sum_{k=1}^{20} X_k \cos\left[\frac{\pi_i \times (k - 0.5)}{20}\right] \qquad i = 1,2,...P \qquad (1)$$

where, P is the order, k is the number of DFT Magnitude coefficients, X_k is the kth order log-energy output from the Mel filter bank.

Thirteen features are extracted for each frame and sampled by the ANN for category classification. The output of the neural network is a classification of each input frame, into their respective categories. By generating a set of output categories for a set of input frames, these categories can be used to match to the target word categories and thus identify the spoken word.

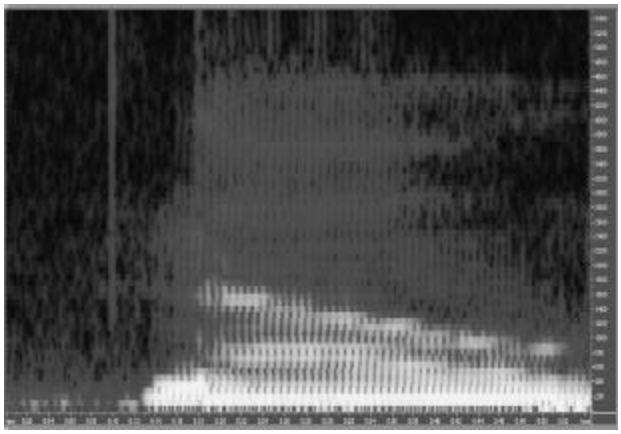

Figure 2. Spectrogram of the spoken word "NO"

Context dependent Category generation

A word such as "yes" would have three broad phonemes: /y/, /E/, and /s/. But because phonemes have a high level of influence on neighboring phonemes, a contextual classification of categories is required [7]. Hence we split each phoneme into one, two, or three parts, depending on the typical duration of that phoneme as well as how much that phoneme will be influenced by surrounding phonemes. As a result, /E/ is split into three parts and /n/ into two parts. The model for the /E/ in "yes" then becomes front<E (/E/ in the context of a preceding front vowel), <E> (/E/ in the middle with no contextual effects), and E>fric (/E/ in the context of a following fricative). Each of these categories is then arbitrarily numbered 1,2, 3 etc so that a target file may be generated, (for training the network), and each category is uniquely mapped to a number. After obtaining the MFCCs and energy features of each speech frame these features are placed in a vector, one vector for each frame, and input to the neural network.

Artificial Neural Network (ANN)

Two ANNs have been trained one for recognizing "yes" and one for "no". These networks have the same architecture and learning algorithm governing their operation.

Architecture of network

The Architecture of the network is the feed-forward structure [7]. The input layer has thirteen processing elements and receives the vector of MFCC and energy values per frame. Then follows one hidden layer of 25 elements. The final layer (outer layer) has 13 output neurons, with each output processing element (neuron) uniquely representing one category.

Algorithm

The training algorithm used for the network is the backpropagation with momentum algorithm. The backpropagation algorithm is applied to feed-forward networks. At each timestep n, let the weight connecting neuron i and neuron j be $w_{ij}(n)$. Let the output from neuron j be y_j. Then we can express the error out of neuron j at time n as:

$$e_j(n) = d_j(n) - y_j(n) \qquad (2)$$

where e is the error, d is the desired output and y is the actual output. Using this error, connection weights are increased in proportion to the error times.

$$W_{ij}(n+1) = W_{ij}(n) + \nu \delta_i(n) y_i(n) \qquad (3)$$

where η is a positive real number less than 1, called the learning rate and δ_j is defined by:

$$\delta_j(n) = e_j(n) \varphi_j'(v_j(n)) \qquad (4)$$

where φ is the transfer function and

$$v_j(n) = \sum_{\forall i} w_{ij}(n) y_i(n) \qquad (5)$$

If j is a hidden neuron, then $d_j(n)$ satisfies the following equation:

$$\delta_j(n) = \varphi(v_j(n)) \sum_{\forall i} \delta_i(n) w_{ij}(n) \qquad (6)$$

Thus $d_j(n)$ is only dependent on the $\delta_i(n)$ in the succeeding layer. [6].

The network was implemented using MATLAB 5.3. Both the ANN units were trained individually, one on "yes" data and another on "no" data. These ANN units were then merged to take in a common input. The outputs of these two networks were then searched for certain patterns using a search function. The network output which best-matched one of the target patterns was then determined to be the "recognized" word.

Pattern Search for Appropriate Sequence

The output of the neural network is a vector, which indicates the category of the input frame. Normally an output vector of a trained network looks like [0.9528 -0.1678 -0.1402 0.0247 0.1139 0.0350 -0.0310 0.0117 -0.0098 0.0314 0.0184 -0.0433 0.0233]. Only one element's value is near 1, and others' are either near zero or have negative values. The maximum element is considered to represent the most probable phonetic category. For example, the vector above has maximum value of 0.9528 in first place, so we can say this frame belongs to category 1 (which is silence). An example of the ANN output is shown in figure 3

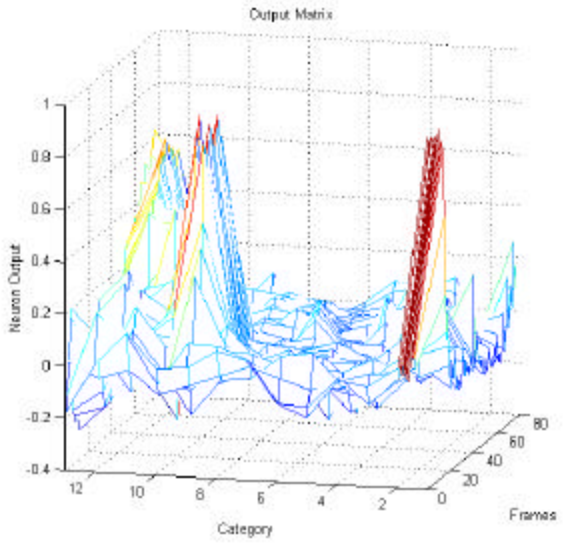

Figure 3. Sample ANN output for "NO"

After feeding the network with the whole word, a matrix of output vectors is generated. The maximum value for each frame output vector is identified and this output is then marked with its appropriate category. A category vector (plot shown in figure 4) is calculated to store the categories of all the frames.

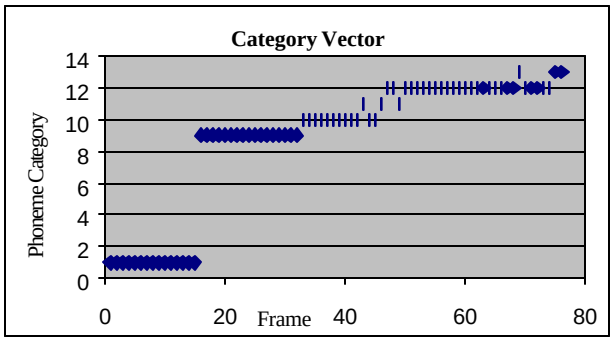

Figure 4. Plot of category vector for "NO" comprised of 76 elements.

The standard output sequence for "Yes" is [2 3 4 5 6 7 8], for "No" is [9 10 11 12 13], excluding silence 1. After attaining this category vector, a refine/search algorithm is implemented on the category vector to conform it to one of the above two sequences. To find the sequence a predefined a transition matrix is used, which describes how one category should transition to another: This transition matrix is shown in figure5.

0	0.5	0	0	0	0	0	0	0.5	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0	0	0	0	0

Figure 5: Sequence Transition Matrix

In this matrix, the [i, j]th element determine if i-th category can transit to j-th category or not. Based on this transition matrix, the category vector is refined using the following steps.

- Remove the replicated numbers. After this step, the vector matrix contains only unique category numbers. For instance, the vector for no_1.wav becomes [1 9 10 11 12 11 12 13 12 13].
- Find the start of the sequence. A search is performed from the element which first has the value 1, and look for the next element (suppose jth element) which will make [1, j]-th element in transition matrix be non-zero. After this step the vector should be [9 10 11 12 11 12 13 12 13].
- Search from the first element of the last element. If the neighboring elements represent an acceptable transition it is recorded. The final output vector should come out as [9 10 11 12 13].

III. TRAINING/TESTING AND RESULTS

The network was trained on a training set of Subject 1's spoken "yes" and "no". The target file was created by assigning a particular category number for a fixed amount of time (determination of the amount of time was done by studying the spectral content of the word) in the file. Thus the first set of frames (in a given set of frames) were assigned to belong to one category and the next set to another category. For example 9 is the desired output for frames/samples obtained between 86 and 190 msecs in figure 2. Thus supervised learning is imparted to the network. Both sub-nets were trained with same scheme, using Subject 1's training voice samples. The maximum epochs were set to 250, and the goal was set to 0.01. After training the network, subject 1's testing voice sample (different from training set) as well as 2 other subject's "yes" and "no" voice samples were tested. The following results were obtained:

TABLE 1
Recognition accuracy of Network

Subject	Number of test Samples	Percentage recognized "Yes"	"No"
1(male)	25	100%	100%
2(male)	20	75%	90%
3(female)	20	50%	60%

Figure 6: Results

IV. DISCUSSION

The system provided satisfactory results. It is robust enough to account for a speaker independent input. Though the encouraging success of the current system is achieved based on a limited vocabulary, the system can be expanded to a larger vocabulary by extending the number of subnets used in the architecture. There was no apparent improvement on the efficiency of the networks on sub-phoneme basis compared to whole word input in this study, subphoneme-level input might significantly reduce the number of networks in regard to the growing vocabulary. Each category is to train on one network correspondingly. The search pattern designed for the current setup is limited for two-word pattern. A more generalized, powerful searching pattern such as Viterbi search and Stack decoding, might be introduced to the architecture. Although the performance of the design is petty satisfactory at current level (100% for subject 1 on whose data the network was trained on and nearly 100% for "no" and 50% above for "yes" of the other subjects), the rate might drop fairly as increasing vocabulary size. The key solution is to increase the number of features extracted on each frame yet at the cost of additional processing time.

V. CONCLUSION

In conclusion the network performs optimally in speaker dependent and limited vocabulary contexts. This network has demonstrated the opportunities and limitation offered by ANN systems. This system can be further expanded to a robust continuous speech speaker independent system when we expand the network and increase training parameters, along with training samples (with sufficient variability). The ultimate goal of this study is to establish an accurate speech recognition system, which is capable of dealing with distinct speech inputs, such as distorted speech (Dysarthric Speech), stressed or elongated speech, rather than just normal speech. Upon preliminary attempts, this network seems to be robust enough to satisfy that task. Further work towards these goals is currently being pursued.

ACKNOWLEDGMENT

We would like to express our thanks to professor Gerald E. Miller for his support. We would also like to thank the CSLU Speech group at the Oregon Graduate Institute of Science and Technology and Mike Brookes in Imperial College whose expertise proved valuable in this research.

REFERENCES

[1] B. Gold and N. Morgan, "Speech and Audio Signal Processing" *John Wiley & Sons, Inc*, 2000.

[2] R. D. Rodman, "Computer Speech Technology" *Artech House Publishers*, pp. 126-128, 1999

[3] M. R. Ashouri "Isolated word recognition using high-order statistics and time-delay neural networks," in *IEEE Signal Processing Workshop on Higher-Order Statistics*, IEEE, Inc., 1997.

[4] C. P. Lim, S. C. Woo, A. S. Loh and R. Osman, "Speech Recognition Using Artificial Neural Networks," *Web Information Systems Engineering*, IEEE, Inc. 2000.

[5] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," *IEEE Trans. J. Magn. Japa*n, vol. 2, pp. 740-741, August 1987 *[Digests 9th Annual Conf. Magnetics Japan*, p. 301, 1982.

[6] S. Haykin, "Neural Networks – *A comprehensive foundation - 2nd ed." Prentice hall*, pp. 156-248,1999

[7] J.-P. Hosom, R. Cole, M. Fanty, "Speech recognition Using Neural networks", CLSU, *Oregon Graduate Institute of Science and Technology*,1999. *(http//cslu.cse.ogi.edu),(http//cslu.cse.ogi.edu/tutordemos)*